SCOTS: Scottish Corpus of Texts and Speech

Jean Anderson, Dave Beavan, Christian Kay

University of Glasgow

1. Introduction

Scotland contains a rich variety of languages. A recent survey of 300 respondents revealed that over 30 different languages are regularly spoken at home and at work (Institute for the Languages of Scotland (ILS) Report 2003). In addition to the Scottish English spoken by the majority, there are substantial numbers of speakers of the other indigenous languages, with an estimated 58,650 Scottish Gaelic speakers at the 2001 census, and about half the population claiming knowledge of Scots (Macafee 2000).[1] Non-indigenous languages include Arabic, Bengali, Cantonese, Dutch, Hindi, Italian, Kurdish, Polish, Panjabi, Romany and Urdu. British Sign Language has a flourishing Scottish variety.

The Scottish Corpus of Texts and Speech (SCOTS) was set up in 2001 to begin the task of making a corpus to represent and monitor the languages of Scotland.[2] Such a corpus is long overdue. Major British corpora, such as the British National Corpus (BNC) and the Bank of English (BE), contain Scottish material but did not collect it comprehensively. The International Corpus of English (ICE) could not find a partner to deal with Scotland in its collection of major varieties of World English. The SCOTS project will thus supply a gap in research materials. Initially we are concentrating on the two most accessible varieties, Scots and Scottish English, but our long-term intention is to sample the languages of Scotland in their totality, thus providing

a snapshot of the linguistic situation in one particular geographical area. Our primary interest is sociolinguistic, matching linguistic patterns to social and demographic categories. Our current chronological cut-off point is 1940, though earlier materials may be included if they are of special interest or contribute to filling gaps. Our informants are not limited to native speakers, since anyone who has lived in Scotland for a substantial period of time may well have been influenced by Scots or Scottish English. Information on place of birth and residence are available in the corpus metadata.

Even our starting point presents the analyst with problems. Information is lacking on how extensively and in what contexts Broad Scots is used, while the range of features characterising Scottish English is not fully defined. Indeed, many of its speakers are unaware that their usage differs in anything but accent from that of speakers of so-called Standard English. Speakers employ features of Broad Scots and Scottish Standard English to different degrees, often depending on context, so that, rather than regarding them as two distinct varieties, it is more accurate to talk about a linguistic continuum running from Scottish English at one end to Broad Scots at the other.[3] At the Scots end there is the further complication of considerable regional diversity, with, for example, marked differences in the speech of the northeast, Glasgow or the Northern Isles. Such diversity is compounded by the fact that there is no agreed spelling system for transcribing these variations. These and other problems will be discussed as the paper progresses.

2. Data Collection

SCOTS was launched in the wake of the devolution of political power from the British Government in London to Scotland.[4] This situation, and especially the re-opening of the Scottish Parliament after a gap of nearly three hundred years, markedly increased public interest in the language and culture of the country. The initial response to our call for data was swift and positive. Offers came from 113 people, and 355 units were identified as suitable. These were processed along with other texts and added to the system to provide a test corpus of around 400,000 words. They included written, spoken and visual materials from a range of genres such as conversation, interviews, correspondence, poetry, fiction and prose. A sub-project was started to collect spoken, written and visual data from the Scottish Parliamentary elections in May 2003. This exercise was repeated for the European elections in June 2004. The materials overall show an expected imbalance, with much of the Scots material being poetry or literary prose (see further 7. below). Each document in the database is accompanied by an extensive set of searchable metadata comprising textual and demographic information and indicating whether the necessary permissions have been received to enable the text to be made public. Such information is currently collected on paper forms, but we expect to develop downloadable versions. The sociolinguistic and administrative data together occupy over 500 fields in the database.

2.1 Types of data

All documents in the corpus are of one or more of the following types:

Text:	document originally conceived as written work
Audio footage:	recording of live speech

Audio transcription:	transcription of speech
Video footage:	recording of live speech with visuals
Video transcription:	transcription of video

In addition we have comprehensive information about the individuals involved:

Author:	author of a document
Participant:	person appearing in audio or video

We also hold associated administrative information, used, for example, in tracking documents from initial contact with the contributor to full copyright clearance, including clearance of third party copyright.

An example of complicated document structure is given below. This scenario is based upon a lecture to a group of people and clearly demonstrates the complex requirements of even a common type of document. The example shows the data collected and the way in which the different elements of the document are mapped to the database under the four major categories described above. Considerable discussion and comparison with other projects went into the development of the sub-categories, but they are not immutable and may be refined further.

The original lecture notes prepared in advance of the talk are identified as a Text, that is a written document. This necessitates compiling the following information:

Text	▪ Text audience (e.g. age, gender, number of people) ▪ Text details (e.g. date by decade, mode of composition)

- Text medium (method of transmission, e.g. book, email)
- Text performance/broadcast details
- Text publication details
- Text setting (domains such as Education, Journalism)
- Text type, with a large number of sub-types:
    - Advertisement (e.g. junk mail)
    - Announcement (e.g. notice)
    - Article (in newspaper, journal, etc.)
    - Correspondence/letters
    - Diary
    - Essay
    - Instructions (e.g. manual, recipe)
    - Invoice/bill/receipt
    - Novel
    - Poem/song/ballad
    - Prepared text (e.g. lecture/talk, sermon, public address/speech)
    - Prose: fiction
    - Prose: non-fiction
    - Report
    - Review
    - Script (film, play, radio, tv etc.)
    - Short story
    - Written record of speech (e.g. Hansard, legal proceedings,

minutes of meetings)

Video footage

Since a recording of the lecture was made, the data are also classified as video footage, requiring information on:

- Audience (e.g. age, gender, number of people)
- Awareness & spontaneity (Did the participants know they were being recorded or might be recorded? Was the event scripted to any extent?)
- Footage information (date, person recording, etc.)
- Footage publication details
- Footage series/collection information (i.e. if it is part of a series)
- Medium (e.g. cinema, radio, telephone)
- Setting (domains such as Education, Journalism plus location, e.g. in a classroom, in Inverness)
- Relationship between recorder/interviewer and speakers (family, friend, professional, etc.)
- Speaker relationships (Did they know one another prior to the recording, and if so, how?)
- There is a further range of types appropriate to this medium:
    - Advertisement
    - Announcement (e.g. news)
    - Commentary
    - Consultation (e.g. medical, legal, business)

- Conversation
- Debate/discussion
- Documentary
- Dramatic performance
- Instruction (e.g. demonstration)
- Interview
- Lecture/talk, sermon, public address/speech
- Lesson/seminar
- Meeting
- Poetry reading/song/ballad performance
- Press-conference
- Prose reading
- Story telling

Video transcription

A transcription was made of the above recording, which necessitates video transcription details:

- transcription publication details
- transcription information
  - Title of original
  - Transcriber identity number in database
  - Conventions
  - Year of transcription

- Year material recorded
- Word count

Author

We also require details of the author of the original text:

- Author details
  - Author identity number in database
  - Name
  - Gender
  - Decade of birth
  - Highest educational attainment
  - Age left school
  - Upbringing (cultural/religious affiliation either now or in the past)
  - Occupation
  - Place of birth
  - Region of birth
  - Birthplace CSD dialect area (=the dialect areas from the *Concise Scots Dictionary* listing)[5]
  - Country of birth
  - Place of residence
  - Region of residence
  - Residence CSD dialect area

- Country of residence
- Father occupation
- Father place of birth
- Father region of birth
- Father birthplace CSD dialect area
- Father country of birth
- Mother occupation
- Mother place of birth
- Mother region of birth
- Mother birthplace CSD dialect area
- Mother country of birth

- Languages
  - Name of language known and whether:
  - Spoken
  - Read
  - Written
  - Understood
  - Circumstances where language is used (e.g. at home or work). A check-list of languages is given, but the contributor may list others.

Participant(s) Similar details are collected for participants, e.g. someone taking part in a discussion, introducing the speaker or presenting a talk of which s/he is

not the author.

As can be seen, describing a document is not necessarily as straightforward as may at first appear since any of these sub-types can in theory be applied in ~~as~~ many combinations ~~as can occur~~. Using the *Concise Scots Dictionary* (CSD) dialect areas enables us to connect our information to a generally recognised classification. All documents, regardless of their type, are treated and manipulated in the same way inside the administrative database (see 4. below). We also have a series of forms to ensure that data protection and copyright legislation are observed. These involved much consultation with the University lawyers and the team learned a great deal as a result.[6]

3. Formats

3.1 Text

We accept document submissions in as many formats as feasible, with a preference for electronic formats wherever possible to reduce processing time. The corpus administration system accepts documents in plain text, maintaining sentence and paragraph breaks only. Plain text was chosen as, for linguistic research, the document content is more important than its presentation; we are also restricted by time and resources. Where a submission is handwritten we take a digital copy of the page. Although we do not currently allow this to be accessed directly by end users, access is allowed by request; we hope this may prove useful to future researchers. The handwritten document is then keyed following strict guidelines and proofread. For typed or printed documents we use optical character recognition (OCR) software to generate text. This method is normally

faster than re-keying, but the lack of a Scots dictionary for the software means that it can generate wrong presumptions about words and the result requires careful proof-reading. For example, scanning of a page of text from a story written in the Doric (northeastern) variety produced *bait* for *hait* 'hot', *oat* for *oot* 'out', and *0* (zero) for *o'* 'of'. The Scots past tense verb ending *–it* was separated from its stem, producing *pump it* rather than the correct *pumpit* 'pumped'. We have also found that writers of Scots texts are not always consistent in their spelling, even within a single text.

3.2 Audio and transcription

When we record data ourselves, we use digital audio tape (DAT) wherever possible. Given the diversity of data we are offered, we must also have the capability to accept source recordings from a number of consumer formats such as tape, CD, Minidisc and computer files. These are immediately converted onto DAT. As it is a high quality digital format we can duplicate and edit DAT material with no quality loss. DAT tapes are thus digitised into the computer with no degradation of quality. Once the information is in computer format we can easily provide members of the team, such as transcribers, with copies of the recording in their format of choice, e.g. CD or Minidisc. The recording must also be edited before its inclusion in the corpus. Typical tasks include trimming the segment (to neaten up the beginning and end of the piece) and noise reduction to combat hiss and background noise.

In the early days of the project, we investigated ways of making these recordings available on the internet. The principal goals were to support a broad user base and to provide flexibility in our own computing requirements. Apple QuickTime was chosen, which gives both PC and Mac users access to the recordings. The server software is open source and free, running

on Linux, Unix or Mac. QuickTime will also allow for streaming of the resource – the end user does not have to wait for the entire piece to download, and can easily jump to any point in the source for near-instant playback. To enable everyone to get the most out of the resource we chose to implement two profiles: one for low bandwidth users (e.g. modem) and a higher quality version for high bandwidth users (e.g. broadband, academic institutions). Transcription is orthographic only, but we aim to provide high enough quality to enable phonetic or phonemic transcriptions to be made by individual researchers.

The transcription guidelines are being refined as our experience grows. Broad Scots raises interesting transcription issues as a language with a range of spoken dialects, and a substantial written record dating back to the 14th century, but no accepted standard written form. This presents problems for corpus building, notably in transcribing spoken data, in identifying specifically Scottish forms, and in lemmatising variant lexical and grammatical forms under search-words. In tackling these issues, we are working with Scottish Language Dictionaries (SLD), which is responsible for the two major Scots dictionaries, the *Scottish National Dictionary* (SND) and the *Dictionary of the Older Scottish Tongue* (DOST), which together form the digitised *Dictionary of the Scots Language* (http://www.dsl.ac.uk/). They supply us with headword listings and are developing a spelling system based on frequency of occurrence of forms.[7]

Examples of transcribed material can be seen in Figure 3 below and on our website. The current guidelines for SCOTS read:

> The *Scottish Corpus of Texts and Speech* is intended to be of use to as wide a range of disciplines as possible, for example to lexicographers, grammarians of Scots/Scottish English, authors and linguists, teachers and pupils. For this reason, the initial

unparsed/unannotated orthographic transcription of spoken language should be flexible, in order to encompass all foreseeable uses of the corpus. Conventional orthography can obscure the presence of spoken language phenomena. This problem is encountered where corpora of spoken English are concerned, but is even more of a concern where spoken Scots and Scottish English is to be adequately represented. Many of the definitive features of Scots and Scottish English are not preserved in English orthography.

Since no phonemic transcription of spoken material collected for the *Scottish Corpus of Texts and Speech* is planned at this stage, it is desirable that the initial orthographic transcription reflects certain important aspects of Scots and Scottish English pronunciation as closely as possible. Scots words should be transcribed using the orthographic representation of each word, as it is found in the *Scottish National Dictionary*. Where there are several options for the spelling of a Scots word, the form that is closest to the pronunciation used by the informant should be selected. At this stage, no Scots spelling conventions have been decided upon, so spellings might have to be normalised at a later date.

The primary orthographic transcription is intended to represent Scots words and Scots/Scottish English forms of pronunciation as closely as possible, by adapting Standard English (henceforth SE) orthography and by making use of existing Scots spelling systems. The transcriber should try to stick to standard English orthography, except where the transcription guidelines indicate otherwise, or where they feel it would enrich the corpus to distinguish between the Scottish pronunciation and the standard English orthographic

representation of the word. Examples might be <a> for SE <all> and <doun/doon> for SE <down>, etc.

3.3 Video

The project team has access to a digital video (DV) camera which we use for all our own data collection. If submissions are of other formats (such as VHS) we use Glasgow University Media Services to assist us in digitisation. Following the same procedures as with audio documents, we initially digitise the source. With DV recordings this is done at zero quality loss as the source is already digital. The computer file is not preserved past the end of processing because of the major storage requirements of the format (2.1 Gigabytes per 10 minutes of footage). Certain editing tasks are performed to neaten up the recording, such as noise reduction and enhancements to the visual presentation. When we have attained the best results, we create an intermediate file for the video compression software to use as its source.

Again as with the audio footage, we use Apple QuickTime to compress the video to suitable sizes for modem and broadband users. The quality is much reduced compared to the original, but this is necessary for delivery to users over the internet. QuickTime provides streaming access so users can jump to any point in the file for playback.

4. Administration system

Our requirements were for a system that could give access to the entire dataset, including the document contents, from one interface. Tight controls on validation and other rules regarding the integrity of the data must be possible. The system must allow access and updates to the data from

different people, at different places, possibly at the same time. For the future, when we begin to collect data from non-indigenous minority languages, the system must be capable of supporting more than just the Latin character set. In addition to these requirements, it must be capable of integrating with other administrative functions such as mail merges, report writing, generation of form letters, and so on. The user interface we provide to all project staff must be easy to use. This is particularly important to reduce training time for casual staff, such as students assisting with data entry.

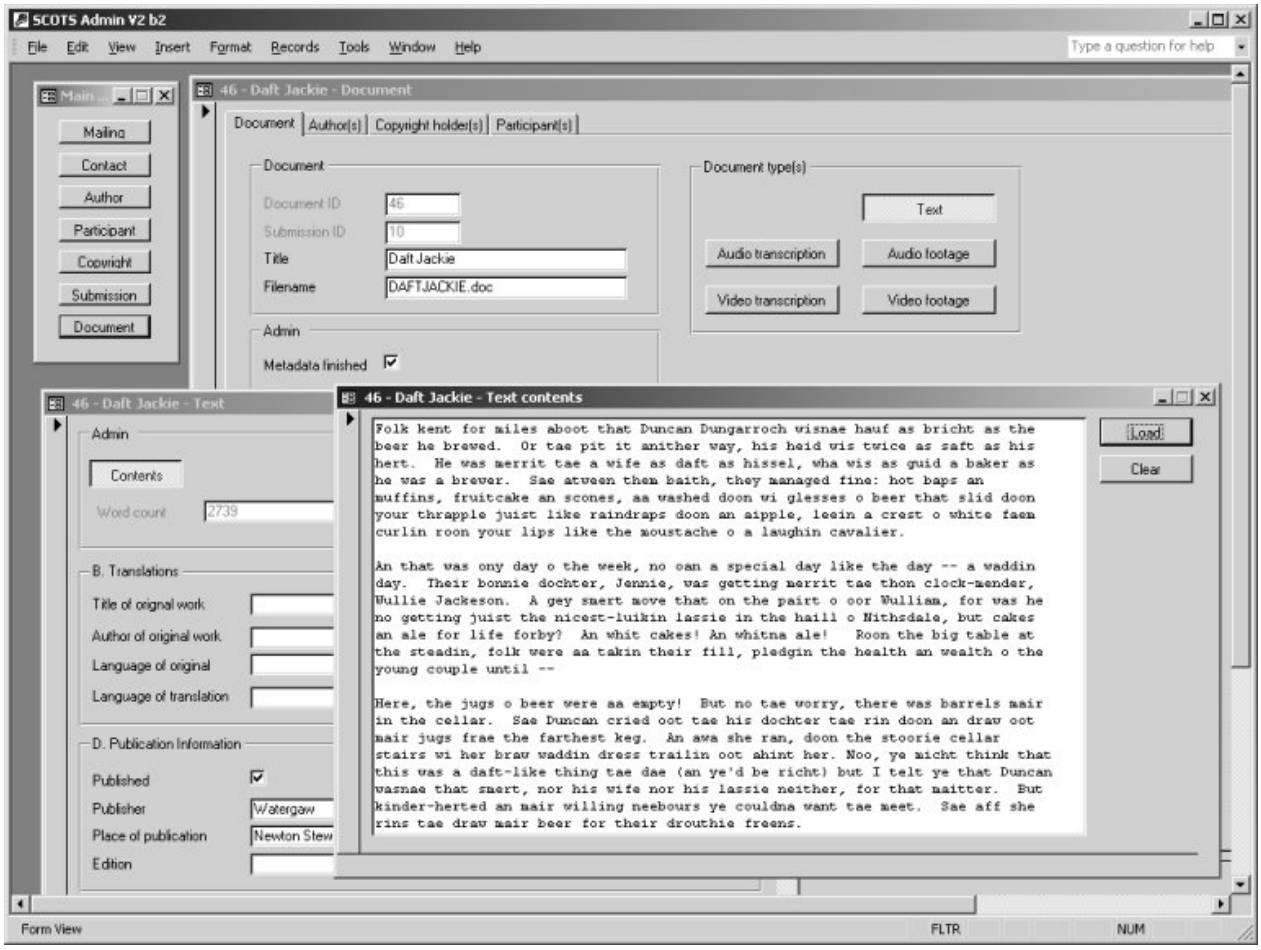

Figure 1. Administration System in MS Access

For storing the raw data a relational database product suits the project very well. To give us the greatest degree of flexibility in the future, and in order not to tie us down to any particular computing platform, an open source, free solution was preferred. After trialling potentially suitable products we chose PostgreSQL. The user interface is implemented in Microsoft Access, which provides an easy transition for users of other Windows software. Reports can be generated and ad-hoc queries made by any user via a visual interface. Integration with the rest of the Microsoft Office suite provides a mail merge facility in Word for numerous correspondence tasks.

There are comprehensive tracking options to record the status of each submission and document. These can be used to ensure we have the correct permissions before public release of the document. Status reports can be generated so that we can identify situations where more investigation or follow-ups are required, for example when a contributor fails to return the necessary metadata forms within a specified period of time.

5. Web site

The sole deployment of the project is our web site. The conditions of our initial EPSRC grant, and now the AHRB grant, require that access be completely free of charge to all users. We expect these users to have a wide variety of interests and to include schoolchildren and teachers, businesses and the general public as well as academics.

In addition to a browse facility, a two-tier search system is needed. The basic first-tier search is in place at the time of writing, offering the facility to interrogate the corpus by a small number of commonly used criteria, for example word/phrase and certain metadata fields such as

age, gender, birthplace, document type, as illustrated for John Corbett's talk, *The Stalking Cure*. Figure 2 shows an initial query and Figure 3 a successful hit.

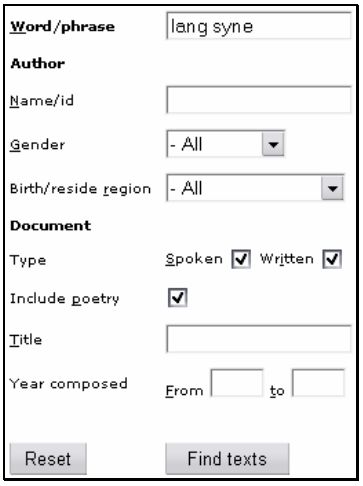

Figure 2. Basic Search Options

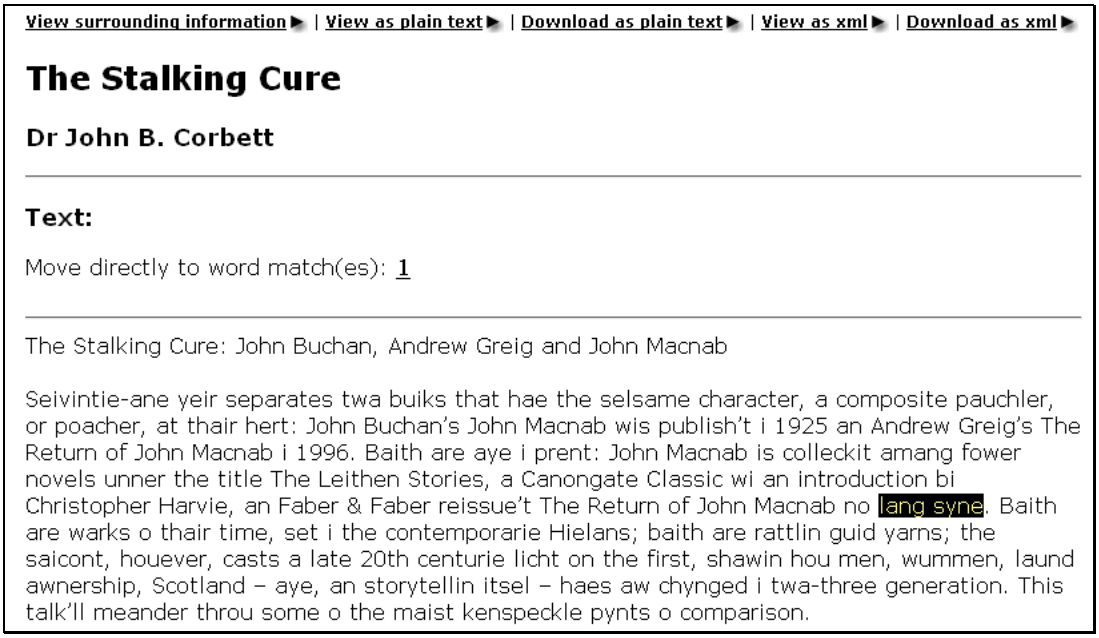

Figure 3. Search result with phrase highlighted

A far more advanced search facility will be developed, to include the facility to use any number of the metadata fields concerning the document, author and participants, as shown in Figure 4. A visual query builder style interface will allow users to combine their criteria together before submitting their search. A list of documents matching the criteria is displayed so that users may select those they wish to view in more detail.

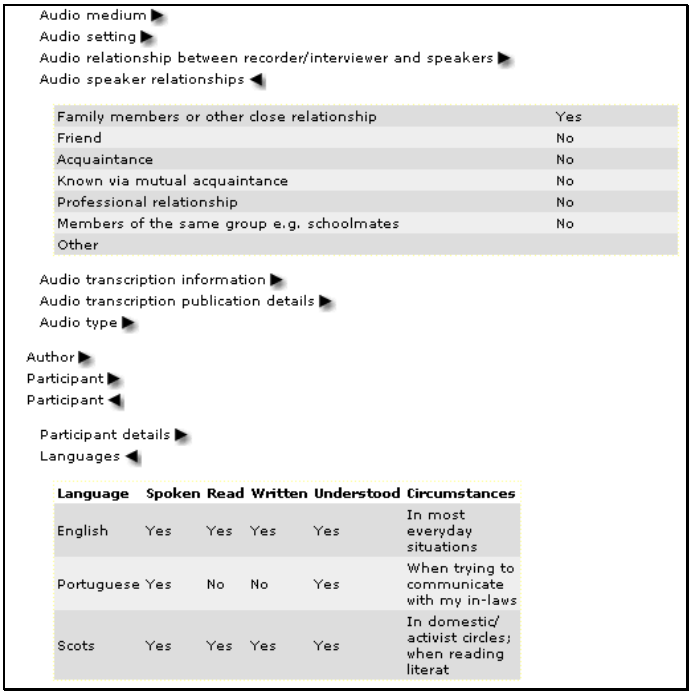

Figure 4. Viewing surrounding information (metadata)

For designing the website we have used Extensible HyperText Markup Language (XHTML) and Cascading Style Sheets (CSS) to provide the greatest accessibility, not only to browsers but also other user agents such as screen readers. For the visually impaired we have a 'high contrast' option which disables any unnecessary graphics and enhances clarity and text size. Documents are viewed by default directly inside the website as HyperText Markup Language (HTML). Provision is also made for viewing and downloading the plain text. In the future, Text Encoding Initiative (TEI) compliant Extensible Markup Language (XML) datasets will also be available. If a search was performed using a word or phrase as a criterion this is highlighted, along with a list to jump quickly to each occurrence of the word or phrase in the document. Access to 'surrounding information' or metadata is given for each document. The different categories as described in 2.1. above can be expanded and contracted at will to reveal the information that the user requires. A list of all recently viewed documents is kept at the bottom of the page to provide quick and easy return access.

We use a relational database (again PostgreSQL) to provide the storage and search facility for the website. All dynamic pages are constructed using templates and processed using PHP scripts. PostgreSQL enables us to make use of built-in advanced text indexing, and we also have the capacity to extend or modify the way this facility works, possibly using word stemming, etc. For security reasons, and to allow more flexibility in the future, the online database is separate from the administrative one. The online database holds only publicly accessible information, which means that a potential security exploit would not release private information.

On a scheduled basis all documents that are marked ready for public use are exported into the online database. At this stage the database ensures that there are no outstanding copyright or IPR issues relating to the document. All administrative information that is not relevant to the document itself or the authors is removed. Where an author or participant has decided to restrict

private information to researchers only this information is not copied. Since contributors can give as much or as little information about themselves as they choose when submitting material, they have total control over what is publicly or privately known about them. If they wish public recognition, as most of the creative writers do, they can, of course, be named as authors.

Currently if a search is performed using a word as a criterion, any matching documents have that word highlighted. One of our priorities for the current phase of the project is to extend this facility and offer users an online concordance. We are working on this in collaboration with the Computing Science Department at the University of Oulu, Finland.

6. Preservation and backup

All correspondence is kept, either in digital form (emails and similar things) or paper (documents sent and received from contributors, authors and so on). If at all possible we keep the original copies of the source documents. If contributors request that these be returned, an archive-quality digital image is made for storage. While processing is in progress, all paper documents are held in secure storage and computer files are accessible by project staff only. Once all the documents relating to a particular submission are fully entered into the administrative system, and processing is finalised, the original documents are passed over to University Archive Services for appropriate storage. Researchers or interested parties may contact Archive Services to be granted access to this material. Any interim revisions of the document are removed once processing is completed, leaving us with the original submission, noting any changes, such as replacement of names which might identify a third party, and the corpus-ready version.

All computer files relating to the corpus are stored on a network drive administered by University Arts IT Support; this storage has a nightly backup with archives stored at University

Archive Services. Our dedicated server for the project has a comprehensive backup schedule to University Computing Service, which gives us the added benefits of restoring files from any period in the last few months if the situation requires it. When conducting major systems work on the server we create a snapshot of the system. This necessitates downtime to stabilise the system. However, if we encounter serious problems, we can revert to a previously good state of the system very quickly.

7. Future plans

During Phase 1 of the project we have developed an easy to use web interface, where no client software is required or dataset updates needed by the end user. As web technology, it is open to any computing platform with a web browser. We have comprehensive sociolinguistic metadata available as well as details about the document. This version has been tested by staff and students and proved to be fast and user-friendly.

During Phase 2 of the project we aim to address the interlinked issues of quantity and balance, and some specifically Scottish issues. Our overall aim for the three years of the AHRB grant is to increase the corpus to at least 4 million words (approximately 800 texts), of which at least 20% will be spoken. It is already clear that Scots survives primarily in speech, and that there should therefore be a concentration on collecting oral data. An advanced search facility will be developed so that users can choose what information to extract. We also plan an online concordance (see 5. above) and a bulk download facility to enable downloading in plain text or XML of all documents which match a particular query. We have no immediate plans to tag the text grammatically, although individual users may do so. An exception might be a ‘light’ tagging of diagnostic features of Scottish English, such as modal auxiliary verbs, present/past participle

constructions and the form of past participles. In the longer term, we will consider vertical expansion to historical texts and horizontal expansion to other languages.

Planning the content of the corpus has raised many questions. We are aware of the need for a corpus to be well-balanced and representative. Most well-known corpora are created from predetermined samples to try to ensure this. The selection criteria for the BNC, for instance, are domain, time, and medium, and target proportions were defined for each of these criteria, as shown below:

Domain: The *domain* of a text indicates the kind of writing it contains. 75% of the written texts were to be chosen from *informative* writings: of which roughly equal quantities should be chosen from the fields of applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs. 25% of the written texts were to be *imaginative*, that is, literary and creative works.

Medium: The *medium* of a text indicates the kind of publication in which it occurs. 60% of written texts were to be books. 25% were to be periodicals (newspapers, magazines). Between 5 and 10% should come from other kinds of miscellaneous published material. Between 5 and 10% should come from unpublished written material such as personal letters and diaries, essays and memoranda. A small amount (less than 5%) should come from material written to be spoken.

Time: The time criterion refers to the date of publication of a text. Being a *synchronic* corpus, the BNC should contain texts from roughly the same period. The intention was that no text should date back further than 1975. This condition was relaxed for imaginative works only, a few of which date back to 1964, because of their continued popularity and consequent effect on the language.

> **Classification Criteria:** a large number of other classification features were identified for the texts in the corpus. No fixed proportions were specified for these features... [This includes such things as Sample size (number of words) and extent (start and end points), Topic or subject of the text, Author's name, age, gender, region of origin, and domicile, Target age group and gender, "Level" of writing (a subjective measure of reading difficulty)] http://www.hcu.ox.ac.uk/BNC/what/writ_design.html

We could not follow such a model initially as we simply did not have enough information about where and how our target varieties are used, or, indeed, what they consist of. In other words, at least some of the data has to be collected before many of the questions implicit in the above categories can be answered. Is Scots used in Scottish newspapers, for example, and if so is it evenly spread throughout or restricted to certain article types such as sport or features? Is the use of Broad Scots in fiction largely restricted to dialogue or does it feature in narrative as well? To what extent, if any, are our two varieties used in informative writing? How do we define Scottish English?[8] Questions of this kind will be familiar to anyone developing a corpus for a language which is not used in all domains of life. Now that the initial data are in place, we will begin to address these questions by identifying under-represented document types and determining whether they are obtainable.

As noted in Section 1 above, political devolution has led to increased interest in linguistic matters. One manifestation of this interest is pressure to set up an Institute for the Languages of Scotland as an umbrella organisation for linguistic endeavours of all kinds. In a modern society, a corpus of the language is an essential part of such an undertaking.

References

Anderson, W. (forthcoming September 2005). The SCOTS Corpus: a resource for language contact study. *Studies in Eurolinguistics*, 4.

Corbett, J. & F. Douglas (2003). Scots in the Public Sphere. In Kirk, J. M. & Ó Baoill, D. P. (eds), *Towards our Goals in Broadcasting, the Press, the Performing Arts and the Economy: Minority Languages in Northern Ireland, the Republic of Ireland, and Scotland*, pp. 198-210. Belfast: Queen's University.

Corbett, J., J. D. McClure & J. Stuart-Smith (eds) (2003). *The Edinburgh Companion to Scots*. Edinburgh: Edinburgh University Press.

Douglas, Fiona M. (2002). The role of Scots lexis in Scottish newspapers. *Scottish Language* 21, pp. 1-12.

Douglas, Fiona M. (2003). The Scottish Corpus of Texts and Speech: Problems of corpus design. *Literary and Linguistic Computing* 18, pp. 23-37.

ILS Report = *An Institute for the Languages of Scotland: A Feasibility Study*, University of Edinburgh, October 2003.

Macafee, Caroline (2000). The demography of Scots: the lessons of the Census campaign. *Scottish Language* 19, 1-44.

Murison, David (1977). *The Guid Scots Tongue*. Edinburgh: James Thin/Mercat Press.

Robinson, Mairi, editor-in-chief (1985). *The Concise Scots Dictionary*. Aberdeen: Aberdeen University Press.

Smith, Jeremy (2000). Scots. In Price, G. (ed.), *Languages in Britain and Ireland*, pp. 159-170. Oxford: Blackwell.

Links

Dictionary of the Scots Language: http://www.dsl.ac.uk/

SCOTS Project: http://www.scottishcorpus.ac.uk/

Scottish Language Dictionaries: http://www.sldl.org.uk/

Institute for the Languages of Scotland:

http://www.arts.ed.ac.uk/celtscot/institutelanguagesscotland/

British National Corpus: http://www.hcu.ox.ac.uk/BNC/what/writ_design.html

For a list of Scots Web links, see also:

http://www.arts.gla.ac.uk/SESLL/STELLA/links.htm#Scots

Abbreviations

AHRB: Arts and Humanities Research Board

BE: Bank of English

BNC: British National Corpus

CSD: Concise Scots Dictionary

CSS: cascading style sheets

DAT: digital audio tape

DOST: Dictionary of the Older Scottish Tongue

DSL: Dictionary of the Scots Language

DV: digital video

EPSRC: Engineering and Physical Sciences Research Council

HTML: HyperText Markup Language

ICE: International Corpus of English

ILS: Institute for the Languages of Scotland

IT: Information Technology

PHP: Hypertext Preprocessor

OCR: optical character recognition

SCOTS: Scottish Corpus of Texts and Speech

TEI: Text Encoding Initiative

XHTML: Extensible HyperText Markup Language

XML: Extensible Markup Language

Notes

[1] 'Scots' is notoriously difficult to define. In this paper it is used interchangeably with 'Broad Scots' and refers to varieties which differ substantially from English Standard English in lexis, grammar and pronunciation. The classic example is the Doric variety spoken in the rural northeast of Scotland. 'Scottish Standard English' is the educated variety, used by many people at all times and by others in writing and more formal speech. The spoken form is characterised by a Scottish accent and minor differences in lexis and grammar, such as use of *wee* 'small' and *bonnie* 'pretty', or different usage of modal verbs, such as *may* and *will*. For information on the history and development of Scots, see Murison (1977), Smith (2000).

[2] The project was funded from 2001-3 by Engineering and Physical Sciences Research Council (EPSRC) Grant GR/R32772/01. It now has a three-year grant from the Arts and Humanities

Research Board (AHRB): B/RE/AN9984/APN17387. We are extremely grateful to both of these bodies for their support. The team currently comprises Dr John Corbett (Principal Investigator), Prof Christian Kay, Dr Jane Stuart-Smith, Jean Anderson, Wendy Anderson (Research Assistant) and Dave Beavan (Computing Manager).

[3] On this and related questions, see Anderson (forthcoming), Corbett et al 2003, Douglas 2003.

[4] The first version of the corpus went live on 30 November 2004 and is available at www.scottishcorpus.ac.uk/ There were 118,000 hits from 2900 visitors on the first day, most of them coming via an article on the BBC News website. The six most popular search words were *gallus* 'bold', *muckle* 'big', *glaikit* 'stupid, *canny* 'careful', *dreich* 'dreary' and *sonsie* 'buxom'. That version contained approximately 532,000 words (385 documents). By 20 January 2005 the total had risen to around 600,000 words (425 documents). Our policy is to update whenever we have significant amounts of new data.

[5] The CSD listing uses County names predating the 1975 Local Government Reorganisation, since most of the work for the parent *Scottish National Dictionary* was based on these divisions.

[6] Most of the work on these forms was done by Dr Fiona Douglas, Research Assistant during Phase 1 of the project and now of the University of Leeds.

[7] We are also grateful to the Newcastle Electronic Corpus of Tyneside English (NECTE) for sharing their expertise in this area with us.

[8] For some answers to such questions, see Douglas 2002, and Corbett & Douglas 2003.

5711 words